APPENDIX K

MEMORANDUM

January 12, 2004

TO: WARREN T. SMITH, President
State Board of Education

FROM: DAVID A. STOLIER
Assistant Attorney General

SUBJECT: **State Board of Education's Duty to Determine Whether Assessment System is Sufficiently Reliable and Valid**

I. Question Presented

You have asked me what the State Board of Education (Board) must do to discharge its legal duty under RCW 28A.655.060(3), which provides in part,

> (c) After a determination is made by the state board of education that the high school assessment system has been implemented and that it is sufficiently reliable and valid, successful completion of the high school assessment shall lead to a certificate of mastery. The certificate of mastery shall be obtained by most students at about the age of sixteen, and is evidence that the student has successfully mastered the essential academic learning requirements during his or her educational career. The certificate of mastery shall be required for graduation but shall not be the only requirement for graduation. . . .
>
> RCW 28A.655.060(3) (emphasis added).

II. Analysis

A. Legislative History

The language was enacted in 1993 as part of section 202 of ESHB 1209. It substantially modified the original education reform legislation from the 1992 session. The same section originally read as follows:

> The academic assessment system shall use a variety of methodologies, including performance-based measures, to determine if students have mastered the essential academic learning requirements, and shall lead to a certificate of mastery. The certificate of mastery shall be required for graduation. . . .
>
> Laws of 1992, ch. 141, § 202(5)(c) (SSB 5953) (emphasis added).

Comparing the amended language to the original language highlights two points. First, the Legislature linked the certificate of mastery (COM) to high school graduation in the original 1992 bill. Second, the 1993 legislation consciously inserted a deliberative process by the Board to trigger the link between performance on the assessment and receipt of the COM. There is little legislative history to be found to illuminate the Legislature's intent. However, I think it fair to say that the Board's role was added as a check step. The Legislature sought some assurance that the assessment system was a fair measure of the skills it was asking students to acquire before performance on the assessment would be linked to a high school diploma.

B. Implementation of an Assessment System

The Board's charge is two-fold. First, the Board is asked to determine that a high school assessment system has been implemented. "Assessment system" is defined in statute as

> [A] series of assessments used to determine if students have successfully learned the essential academic learning requirements. The assessment system shall be developed under *RCW 28A.630.885(3)(b).[1]
>
> RCW 28A.655.010(5).

[1] The asterisk denotes that RCW 28A.630.885 was recodified as RCW 28A.655.060.

The referenced statute directs the Commission on Student Learning[2] and the Office of Superintendent of Public Instruction (OSPI) to develop a statewide assessment system for use in the elementary, middle, and high school years designed to determine if each student has learned the essential academic learning requirements (EALRs). The assessment system must include criterion-referenced and performance-based measures. RCW 28A.655.060(3)(b)(i). The system must be designed so that the results under the assessment system may be used by educators as tools to evaluate instructional practices and to initiate appropriate educational support for students who have not learned the EALRs at the appropriate periods in the student's educational development. RCW 28A.655.060(3)(b)(ii).

Therefore, the first part of the Board's charge is to determine if a high school assessment system that meets the above-referenced criteria has been implemented in the state. The Board should bear in mind that it may not have a "finished" assessment system to opine on. The Legislature contemplated that both the EALRs and the assessments would continue to evolve, specifically authorizing OSPI to modify both as needed. RCW 28A.655.060(3)(b)(v).

C. Determination of Sufficient Reliability and Validity

The second charge to the Board is to determine whether the high school assessment system is sufficiently reliable and valid. We first need to define terms. This presents a challenge because there is not universal agreement on the use of the word "validity" and the Legislature did not specify what it meant. I have reviewed enough material to understand that "validity" can be a loaded term and usage varies among experts and policy advocates. In addition to the working definitions previously adopted by the Board, I will introduce definitions used in the two leading court cases reviewing the legality of high stakes tests: GI Forum, Image De Tejas v. Texas Education Agency[3] in Texas, and Debra P. v. Turlington[4] in Florida. The courts used reliability and validity as tools to evaluate whether high stakes assessments were fundamentally fair under the Due Process Clause and/or the Equal Protection Clause of the Constitution.

---

[2] The various powers, duties, and functions of the Commission on Student Learning were divided up between the Superintendent of Public Instruction and the A+ Commission and transferred as of July 1, 1999. RCW 28A.655.900.

[3] 87 F. Supp. 2d 667 (W.D. Tex. 2000).

[4] The Debra P. litigation over Florida's State Student Assessment Test (SSAT) as a graduation requirement resulted in multiple court decisions as it bounced between the federal district court and federal court of appeals over the course of 5 years. I refer to them as follows: Debra P. I, 474 F. Supp. 244 (M.D. Fla. 1979); Debra P. II, 644 F.2d 397 (5th Cir. 1981); Debra P. III, 564 F. Supp. 177 (M.D. Fla. 1983); Debra P. IV, 730 F.2d 1405 (11th Cir. 1984).

1. Reliability. Reliability generally refers to how often a test will yield the same result. It is an indicator of the consistency of measurement. GI Forum, 87 F. Supp. 2d at 672. The Board has been working with the following definition:

> Reliability is the degree to which the results of an assessment are dependable (i.e., relatively free from random errors of measurement) and consistently measure particular student knowledge and/or skills. . . .[5]

These definitions are consistent. The issue of sufficient reliability does not seem to have been a very contentious one in the court cases, leading me to believe that neither the definition nor the methodology for demonstrating reliability are overly controversial. Based on information presented at the most recent Board meeting, it appears that the Board should be able to tap into OSPI staff for statistical evidence of reliability.

2. Validity. Validity is a more difficult term to pin down because there are various types of validity that could apply to competency testing. The Board has been working thus far with the following definition:

> Validity is the extent to which an assessment/test measures what it is supposed to measure, as well as the extent to which inferences and actions based on the assessment/test scores are appropriate and accurate. . . .[6]

According to the Texas court, validity generally refers to the weight of the accumulated evidence supporting the particular use of the test scores. GI Forum, 87 F. Supp. 2d at 672. The Florida courts used the term "content validity" to refer to the degree to which the test measures the knowledge and skills sought to be measured. Debra P. II, 644 F.2d at 404, n.10.

"Construct validity" is a related concept that refers to how well the test measures the construct for which it was designed. Debra P. II, 644 F.2d at 404, n.10. In other words, does the performance on the test really reveal whether the student can comprehend what she reads or can solve math problems? For purposes of criterion-referenced assessment of academic skills, there appears to be little difference

[5] Final Report of the Certificate of Mastery Study Committee, May 2003.

[6] Final Report of the Certificate of Mastery Study Committee, May 2003.

between construct and content validity, since the construct tested is the mastery of the required academic content. Thus, the construct validity is grounded in the content validity of the test.[7]

To scope down another level, courts have also used the terms "curricular validity" and "instructional validity" as subsets or components of content validity. The courts have used the two terms inconsistently, but both have been used to refer to the notion that an assessment must measure material that is taught to students. Another way to put it is that students must have been afforded adequate opportunity to learn the material covered on the assessment.

See GI Forum, 87 F. Supp. 2d at 672; Debra P. II, 644 F.2d at 404.

The Florida court distinguished the two terms in the following way. Curricular validity means the test parallels the curricular goals of the state (i.e., the EALRs in Washington). See Debra P. III, 564 F. Supp. at 184. Instructional validity is an "elusive concept" that ensures the test is a fair test of that which is taught in the schools. Id.; Debra P. IV, 730 F.2d at 1407. Not everyone agrees on the appropriate use of the two terms. For instance, the Texas court did not use the term "instructional validity". However, it did address "opportunity to learn" as a fairness issue.

I believe the Board's working definition captures the core of the concepts of content and construct validity and for reasons set forth below most likely comports with the Legislature's understanding of the term when it enacted the legislation.

3. Sufficiency. The word "sufficient" in the phrase "sufficiently reliable and valid", may be defined as "enough to meet the needs of a situation or a proposed end".[8] It necessarily calls for the Board to exercise judgment and discretion. To do so, the Board needs some sense of what the purpose of the assessment is. Similarly, in order for the Board to apply its working definition of validity, it needs to be mindful of what the assessment is supposed to measure.

---

[7] See William A. Mehrens, "Defending a State Graduation Test: *GI Forum v. Texas Education Agency*. Measurement Perspectives From an External Evaluator", Applied Measurement in Education, 13(4) p. 395 (2001).

[8] Miriam-Webster On-line Dictionary.

4. <u>Purpose of the Assessment</u>

To determine the purpose of the assessment, it is appropriate to look at the full scope of the education reform legislation. There is ample redundant language to suggest where the Legislature's focus was. "The certificate of mastery . . . is evidence that the student has successfully mastered the essential academic learning requirements during his or her educational career. . . ." RCW 28A.655.060(3)(c). The assessment system is "designed to determine if each student has learned the essential academic learning requirements . . . ." RCW 28A.655.060(3)(b)(i). The assessments must be "directly related to the essential academic learning requirements . . . ." RCW 28A.655.060(3)(b)(vi). Finally, recall the "assessment system" is defined as one used to determine if students have successfully learned the EALRs. RCW 28A.655.010(5).

As a whole, the legislation focuses on constructing a system that (1) identifies skills that students should know; (2) develops EALRs based on those skills; and (3) develops assessments to fairly measure mastery of the EALRs. At the same time, the Legislature has embedded instruction of the EALRs into the basic education program, requiring school districts to make the EALRs part of the program offered to all students. RCW 28A.150.220(1)(b).[9] Within this context, the most reasonable interpretation is that the Legislature gave the Board the duty to "close the loop". That is, the Board must assure that the assessment really does measure students' mastery of the EALRs.

Therefore, I conclude that Board's primary duty is to determine whether the high school assessment system is sufficiently reliable and valid for purposes of determining whether students have mastered the EALRs. In other words, the Board must be satisfied that the assessment measures the competencies it is supposed to measure and does so reliably.

Secondarily, I think there is some implied responsibility for the Board to consider and comment on the issue of fairness (or opportunity to learn) raised by its duty to trigger the graduation requirement. Whether the Legislature was aware of the legal implications of high stakes tests or not, the fact remains that the Board's determination does trigger the graduation requirement. According to the case law

[9] This requirement also came in with the original 1992 legislation. Laws of 1992, ch. 141, § 503.

discussed below, even if the assessment reliably and validly measures mastery of the EALRs, fairness dictates that students have actually been exposed to the material before successful performance is required for graduation.

I turn now to a brief discussion of the Texas and Florida cases to illustrate the different ways evidence of validity was used in examining the legal fairness of high stakes tests.

D. High Stakes Tests—Evidence of Validity and Opportunity to Learn

As the Florida and Texas cases demonstrate, test reliability and validity inform fairness issues, such as the opportunity to learn. The cases do not set forth a rigid formula for defending the use of high stakes exams. Rather, in each case some combination of factors provided sufficient weight to satisfy the court.

In the Florida case, the trial court initially held the Florida assessment had sufficient content validity. The appellate court subsequently determined the holding was erroneous because the record lacked evidence that the material covered on the test was actually studied in the classrooms of the state. Debra P. II, 644 F.2d at 404. The appellate court, therefore, remanded the case (sent the case back) to the trial court. Thereafter, the state commissioned a consultant to develop and administer a study, consisting of a variety of surveys. The trial court relied on the study as well as state policies regarding curriculum, retakes, and remediation to conclude that the assessment was instructionally valid and that students had an adequate opportunity to learn. Debra P. III, 564 F. Supp. at 184-86. Specifically, the following factors were taken into consideration by the court as evidence that the assessments were sufficiently valid to provide fundamental fairness to the students,

- Uniform testing standards at various benchmark grades to monitor the acquisition of basic skills by students statewide.

  (Debra P. III, 564 F. Supp. at 185.)

- A four-part survey to determine whether the school districts teach the skills tested by the competency exam: the components were a teacher survey, a district survey, site visits to verify the district reports, and a random student survey.

  (Debra P. III, 564 F. Supp. at 180-82.)

- Pupil progression plans to ensure that students are not promoted without consideration of each student's mastery of basic skills.

(Debra P. III, 564 F. Supp. at 185.)

- Students given multiple chances (5) to pass the test. If they failed, they were offered state-funded remedial help targeted at the student's identified deficiencies. Remediation efforts were monitored by the state.

(Debra P. III, 564 F. Supp. at 185; Debra P. IV, 730 F.2d at 1411.)

- School districts no longer had authority to decide not to teach the minimum standards. The state department of education published and distributed minimum performance standards; the state periodically reviewed programs for compliance; districts provided annual reports to the state; districts had access to state-approved instructional materials.

(Debra P. III, 564 F. Supp. at 184.)

Although it did make use of the survey in its second decision, the trial court also recognized that the issue of instructional validity is a slippery slope. The court took pains to note that it would be impossible to prove conclusively the degree to which every one of the students were exposed to the skills measured on the test.[10] Rather, "[w]hat is required is that the skills be included in the official curriculum and that the majority of the teachers recognize them as being something they should teach. . . ." Debra P. III, 564 F. Supp. at 186. When this has been shown, then "the only logical inference is that the teachers are doing the job they are paid to do and are teaching these skills. . . ." Id. The second appellate court agreed. It specifically rejected the appellants' argument that there must be direct evidence that students were "actually taught" the subjects tested. Debra P. IV, 730 F.2d at 408.

The Texas court took an approach similar to where the Florida courts ended up. In the GI Forum decision, the court focused on construction of the assessment, alignment of the assessment to state's essential skills, and remediation provisions. Specifically, the court relied on the following:

- Rigid state-mandated correlation between the TEKS (Texas version of the EALRs) and the assessment.

(GI Forum, 87 F. Supp. 2d at 674.)

[10] "[A]bsent viewing a videotape of every student's school career, how can we know what really happened to each child? . . ." Debra P. III, 564 F. Supp. at 184.

- Testimony by experts of the actual assessment and item development process, including piloting and review processes.

(GI Forum, 87 F. Supp. 2d at 672.)

- Reviews of test items during test construction for whether the items covered sufficiently-taught portions of the state-mandated curriculum.

(GI Forum, 87 F. Supp. 2d at 672.)

- State-mandated remediation on specific subject areas. Even though there was no state-mandated approach to remediation, the state was able to produce evidence of successful remediation.

(GI Forum, 87 F. Supp. 2d at 673.)

- Eight opportunities for students to pass the exam prior to their scheduled graduation date, meaning a single test score did not serve as the sole criterion for graduation.

(GI Forum, 87 F. Supp. 2d at 675.)

The Texas court put this evidence together and concluded that since (a) the assessment measures what it purports to measure; (b) does so with a sufficient degree of reliability; and (c) the state has made largely successful efforts at remediation and offered substantial retake opportunities, all students had a reasonable opportunity to learn. GI Forum, 87 F. Supp. 2d at 682.

The courts in both cases put great weight on remediation efforts. The Texas court noted that the result of poor performance on the exam was additional, targeted educational opportunity for students and another chance at passing the test. GI Forum, 87 F. Supp. 2d at 674. The Florida courts similarly mentioned that remedial instruction targeted at students' identified deficiencies bolstered a finding of instructional validity. Debra P. IV, 730 F.2d at 1408, 1410. Thus, the opportunity for retakes and targeted remediation may substantially cure systemic deficiencies that otherwise weigh against finding an adequate opportunity to learn.

E. Implications of the Court Decisions for This Board's Role

In contrast to the role of the courts, the Board in Washington has not been asked to undertake a constitutional review. Rather, it has been asked to certify that a necessary component of fundamental fairness is in place: that is, whether the assessment sufficiently measures what it purports to measure and does so with a sufficient degree of reliability.

Nonetheless, I believe there are some important lessons to be taken from the cases. First, taken together the cases demonstrate how slippery the term "validity" can be. The one common denominator is that it is essential that an assessment measure the basic skills it purports to measure. Therefore, even a conservative definition of validity serves broader fairness concerns. Second, the cases identify additional components of fairness that will come into play once the COM becomes a graduation requirement. The court cases teach that the State should have in place several structural components as indicia of opportunity to learn. These include the EALRs being established as part of a state-mandated curriculum; schools and teachers having access to the EALRs; and students having retake and remediation opportunities. These clearly will be critical issues once the COM becomes a graduation requirement. To the extent the Board perceives any of these components are missing, it should advise the Legislature.

The Board need not peer into each classroom of the state. The cases cast doubt on whether such an exercise is legally necessary or even possible. Further, the Legislature did not intend the Board to engage in that level of scrutiny when it gave the Board the authority and duty to determine reliability and validity of the assessment system. Rather, the Legislature gave the Board a role to mark the appropriate beginning of the next stage of an ongoing process.[11]

[11] Until the Board triggers the COM, there is no way to evaluate the effect of a diploma sanction on opportunity to learn. The Texas court recognized that "there is a measurable difference in the motivation between students taking a field examination and students taking a test with actual consequences. . . ." GI Forum, 87 F. Supp. 2d at 673. Similarly, although the graduation requirement in Florida was postponed by the litigation, the court thought it likely that the threat of the diploma sanction pending the outcome of the litigation contributed to improved pass rates. Debra P. IV, 730 F.2d at 1416.

III. Conclusion

The Legislature charged the Commission on Student Learning and OSPI to develop the EALRs and develop an assessment designed to measure mastery of the EALRs. It charged school districts to provide a basic education program that includes the EALRs. It charged this Board to determine that the high school assessment system is sufficiently reliable and valid for measuring whether students have mastered the EALRs. If and when the Board makes a positive determination, the COM will become a graduation requirement. In the narrowest sense, the Board will have discharged its legal duty at that point.

The fact that the Board has been placed in the position of triggering the graduation requirement implies that the Board should also be aware of the "fairness" issues. Case law regarding the fairness of high stakes tests suggests that the most critical components of fundamental fairness are the following: (1) sufficient reliability and validity of the test, i.e., whether the test measures what it purports to measure and does so with a sufficient degree of reliability; (2) a requirement that the measured skills be taught; (3) sufficient notice that successful performance on the assessment will be required for graduation; (4) opportunities for students to retake the exam; and (5) remediation opportunities for students who fail to successfully perform on the exam.

Therefore, I believe the Board could and should appropriately advise publicly on the presence or absence of the recognized components of fairness as a corollary to determining reliability and validity of the assessment.

I trust this is of some assistance. This memorandum is not an official Attorney General Opinion, but represents my own considered analysis as your assigned assistant attorney general.

_____

DAVID A. STOLIER
Assistant Attorney General
(360) 586-0279

DAS:bb